An Interlingual-based Approach to Reference Resolution

David Farwell
Computing Research Laboratory
Box 30001/3CRL
New Mexico State University
Las Cruces, New Mexico 88003
david@crl.nmsu.edu

Stephen Helmreich
Computing Research Laboratory
Box 30001/3CRL
New Mexico State University
Las Cruces, New Mexico 88003
shelmrei@crl.nmsu.edu

Abstract

In this paper we outline an interlingual-based procedure for resolving reference and suggest a practical approach to implementing it. We assume a two-stage language analysis system. First, a syntactic analysis of an input text results in a functional structure in which certain cases of pronominal reference are resolved. Second, the f-structure is mapped onto an interlingual representation. As part of this mapping, the reference of the various f-structure elements is resolved resulting in the addition of information to certain existing IL objects (coreference) or in the creation of new IL objects which are added to the domain of discourse (initial reference).

For this effort, we adopt Text Meaning Representation for our IL and rely on the ONTOS ontology (Mahesh & Nirenburg, 1995) as a general knowledge base. Since the central barrier to developing such a system today is the incompleteness of the knowledge base, we outline a strategy starting with the implementation of a series of form-based resolution algorithms that are applied directly to the referring expressions of the input text. These are initially supplemented by a knowledge-based resolution procedure which, as the knowledge base grows and the adequacy of the f-structure and IL-representation increases, takes on more and more of the processing load.

We examine the operation of the form-based algorithms on a sample Spanish text and show their limitations. We then demonstrate how an IL-based approach can be used to resolve the problematic cases of reference. This research effort is part of the CREST project at the CRL funded by DARPA[1].

1 Introduction and Background

This paper describes a practical approach to implementing an interlingual-based reference resolution procedure. This effort is part of a recently initiated research project at the Computing Research Laboratory at New Mexico State University which aims to develop a system that will automatically construct an epistemic knowledge base from text. The procedure will also support machine translation (e.g., Mikrokosmos, see Onyshkevych & Nirenburg, 1994; Carlson & Nirenburg, 1990) as well as other multilingual NLP tasks. Since the system itself is in the early stages of development, descriptions of performance are projected rather than actual.

Current approaches to reference resolution focus on surface forms (see Sundheim & Grishman, 1995, and MUC-7, 1998, for a general introduction). They include string match algorithms for proper noun phrases (PNs) and common noun phrases (NPs) (e.g., Bagga & Baldwin, 1998) and syntactically constrained morphological matching for pronominals and deictic NPs (Lappin & Leass, 1994). In some cases, they include constraints related to clause structure or larger textual units (e.g., Grosz, Joshi, and Weinstein, 1995).

[1] DOD contract N 66001-99-1-8915

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2000**	2. REPORT TYPE	3. DATES COVERED **00-00-2000 to 00-00-2000**

4. TITLE AND SUBTITLE **An Interlingual-based Approach to Reference Resolution**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Computer Science,New Mexico State University,PO Box 30001,Las Cruces,NM,88003-8001**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **11**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

They do not generally consider implicit references or, in the case of Spanish (e.g., Ferrández *et al*, 1998), references to contextually clear possessors using determiners rather than possessive adjectives. All such approaches are supplemented, if not entirely determined, by heuristics which, more recently, have been induced statistically from corpora (e.g., Hirschman *et al*, 1998, Popescu-Belis, 1998). The only approaches of which the authors are aware that attempted to account for implicit referents or implicit references are those developed within AI over two decades ago (e.g., Hobbs, 1979, or DeJong, 1979).

This approach suggested here differs radically in that reference resolution is triggered by elements of interlingual (IL) representation rather than surface text expressions. The referents in the domain of discourse consist of elements of IL representation as well. Thus, implicit references and implicit referents are accounted for and, at the same time, empty references are ignored.

Below, in Section 2, we outline a proposal for practically implementing an IL-based approach, beginning with a description of a target procedure which resolves the reference of each new IL element as it is being produced, clause by clause. We then present a series of form-based procedures, which are to be gradually replaced (or, in some cases, transmogrified) as the IL analysis system and supporting knowledge bases are extended. In Section 3, we briefly describe the relevant aspects of the sample Spanish used as a basis for the presentation. In Section 4, we examine the operation of the procedures in greater detail and demonstrate how the IL-based approach can resolve those problematic references that are beyond the scope of the form-based approaches.

2 Resolution Procedures

2.1 IL-based Resolution Procedure

The proposed approach relies on the capability of a system to provide a reasonably adequate interlingual representation for a text. Here the interlingua we rely on is a variant of Text Meaning Representation (TMR) (see http://crl.nmsu.edu/Research/Projects/mikro/index.html) and we focus on a sample Spanish text and its interlingual analysis, which is known to be reproducible automatically.

The first step of the analysis process is to produce a functional (syntactic) structure for the text. As part of the process of establishing the f-structure, various structurally governed, clause and sentence internal anaphoric relations will be resolved, the relevant anaphors being coindexed (or assigned differing indexes) as determined by the syntax. Thus some syntactic co-reference relationships such as those related to clitics and relative pronouns will be identified before the IL procedure begins.

The second step is to map from the f-structure to the TMR. A TMR includes, among other representational objects, instantiations of object types, relation types and property types. These are constructed from ontological concepts which are associated with the lexical items in f-structure and which are filled out on the basis of the surrounding f-structure.

For instance, the Spanish verb *comprar* (to buy) might be associated with the ontological concept named PURCHASE which is a generic frame structure corresponding to purchasing events. It might in part look like:

```
TIME: T [time]
LOCATION: L [place]
PURCHASE
    AGENT: P [human/organization]
    THEME: O [object]
    SOURCE: S [human/organization]
    AMOUNT: C [currency]
    etc.
```

where TIME, LOCATION, AGENT, THEME, human, organization, object, etc. are all ontological concepts. On some particular use of *comprar* in a text (or more specifically in the f-structure representation of a text), the PURCHASE frame is called forth and instantiated, i.e., indexed and filled in with instantiated representational objects derived from other ontological concepts associated with other lexical items of the surrounding f-structure.

For instance, if the f-structure in question is for:

Roche compra Docteur Andreu

Roche and *Docteur Andreu* will each be associated with and instantiate an ontological concept for COMPANY. Such a representational element might look like:

```
COMPANY
    NAME: N [text]
    HEADQUARTERS: H [office]
    BRANCHES: B [set:office]
    SUBSIDIARIES: S [set:company]
    EMPLOYEES: E [set:human]
    SALES: V [currency]
    PROFITS: P [currency],
    etc.
```

where NAME, HEADQUARTERS, BRANCH, text, office, etc. are all ontological concepts. This frame is instantiated according to the f-structure context for each of the two proper names and then the instantiated frames are inserted (actually indexed) appropriately in the PURCHASE frame. The result is:

```
COMPANY-1
    NAME: Roche [text]
    HEADQUARTERS: [office]
    etc.

COMPANY-2
    NAME: Docteur Andreu [text]
    HEADQUARTERS: [office]
    etc.

TIME: dateline-time [time]
LOCATION: L [place]
PURCHASE-1
    AGENT: COMPANY-1 [human/org]
    THEME: COMPANY-2 [object]
    SOURCE: S [human/org]
    AMOUNT: C [currency]
    etc.
```

These instantiated representational objects are, in turn, referents in the discourse context when the next sentence is processed. As part of this mapping, the reference of the various f-structure elements is resolved resulting in the addition of information to certain existing IL objects (coreference) or in the creation of new IL objects which are added to the domain of discourse (initial reference). Which occurs depends on whether a connection can be inferred between the current IL object and an already-existing IL object on the basis of ontological or epistemic information.

Note that reference resolution is driven by the process of instantiating TMR objects rather than by linguistic forms. At the same time, various aspects of reference resolution are being done on the basis of form. Certain anaphors are coindexed or differently indexed on the basis of morphosyntax as the f-structure is being constructed. Ontological concepts of similar or related type may be called forth during the mapping of similar or related lexical items onto TMR. Articles and other determiners affect the form of the instantiated TMR objects corresponding to the NPs containing them. Finally, aspects of the instantiated TMR objects refer to literals such as the NAME attribute of a COMPANY. These can be used to implement string-matching algorithms.

2.2 Approach to Implementation

Because the target procedure relies on a highly sophisticated, and as yet incomplete, language analysis system, the approach to implementing it begins by assuming that none of the IL apparatus is available for processing. Instead, we implement an initial set of standard, form-based algorithms which vary according to the syntactic category of referring expression.

Generally, if the referring expression is a proper noun phrase, a full or partial string match with each prior PN is used to establish a coreference link. Otherwise, the PN is assumed to refer to a new referent. For standard pronominals, a recency algorithm is used which checks the morphosyntactic constraints (gender, number, etc.) and, if possible, the semantic class to filter potential coreferents. When a match is found, the coreference link is established. Common noun phrases follow a bifurcated algorithm. If the noun phrase is indefinite or has no article, then it is assumed to refer to a new referent. If it is a definite noun phrase, then the head noun string is matched against that of previous NPs. If the heads of two NPs match and the complement strings do not mismatch, the two NPs are assumed to corefer.

As the IL analysis system and supporting knowledge bases grow and the ability to produce appropriate f-structures and TMRs is extended, the target procedures will bear increasingly more of the resolution task.

3 The Data

For this presentation, we focus the discussion on a single Spanish text, a newswire article concerning a corporate buyout (see Appendix I for the original text and its translation into English). It is taken from the ARPA Machine Translation Evaluation corpus (White *et al*, 1994) and contains 347 words, 17 sentences and 11 paragraphs.

There are 144 referring expressions altogether including 14 proper noun phrases (9.7%), 76 common noun phrases (52.8%), 20 pronominal-like expressions (13.9%), 31 verbal expressions (21.5%) and 3 prepositional phrases (2.1%). Of the common noun phrases (NP), 43 were definite NPs, 8 were indefinite NPs and 25 NPs had no explicit determiner. Of the pronominal expressions, 12 were pronouns (Pron) or deictics such as *hoy* (today), *aquí* (here) or *ahora* (now), 3 were ellipted (subject) pronouns (PRO) and 5 were definite articles which function as possessive adjectives (Det [= Pron]) as in:

> *El* *beneficio neto ... se elevó ....*
> The[= **Its**] net profits ... increased ....

However, there are in addition 129 implicit references made which need to be resolved as well. Implicit references are those that are implied by the slots of TMR objects. They may be of unexpressed participants of an event (say, a seller in a PURCHASE event), or unexpressed times or locations of events. Altogether, then, there are 273 references of which 52.75% are explicit and an almost equal amount, 47.25%, are implicit. Taking the implicit references into account, the proper noun phrases (PN) represent 5.1% of the referring expressions, the common noun phrases NP about 27.8%, pronominal-like expressions 7.3%, verbal expressions 11.4% and prepositional phrases about 1.1%. These results are summarized in Tables 1 and 2.

As for referents, there are 138 altogether. Of these 108 (78.25%) are referred to explicitly on at least one occasion while 30 (21.75%) are referred to implicitly only. There are 40 (29%) referents that are referred to more than once, of which 31 (77.5%) are explicitly referred to at some point and 9 (22.5%) are implicitly referred to only.

Total References	Explicit References	Implicit References
273	144	129
100%	52.75%	47.25%

Table 1. Total references

Proper Nouns	Comn Nouns	Prons	Vrbls	Prep Phrases
14	76	20	31	3
9.7%	52.8%	13.9%	21.5%	2.1%

Table 2. Types of Explicit References

Thus, there were 135 coreferences altogether (273 total references – 138 referents), 41 explicit coreferences and 94 implicit coreferences. Of the explicit coreferences, 9 were made by PNs, 10 by NPs, 20 by pronominal-like expressions (12 by Prons, 3 by PROs, 5 by Dets [= Pron]) and 2 by verbal expressions.

It is perhaps of some interest that very few of these references were figurative, that is, metonymic or metaphorical. There was one clearly metonymic reference, and one or two other possible metonymic references.

4 Reference Resolution

What follows here is a discussion of what is needed to resolve the 273 references made in the example Spanish text. In each section, the cases that could be handled by a form-based method are discussed, then cases that would require an IL-based approach for resolution.

4.1 Resolving Proper Noun Phrases

As mentioned, there are 14 PNs in the sample text, 9.7% of the explicit referring expressions or 5.1% of all references. In addition, there were 4 cases of common noun phrases which were in apposition with proper noun phrases and which will be considered here as well.

For PNs, the basic resolution strategy is to match the form of the expression with that of each of the PNs used previously. If there is a match, assume the current PN is being used to corefer to the referent of the matching PN. This applies in 4 of 18 cases. If no match is found, attempt a partial positive form match. That is, if any part of either form for which there is a partial match does not match a corresponding substring in the other form, then the match fails. Thus, *El grupo Roche* matches *Roche* and *Productos Roche SA* also matches *Roche* because there is no substring of the phrase *Roche* that does not match with *El grupo* or *Productos*. On the other hand, *Productos Roche SA* does NOT match *El grupo Roche* because *Productos* does not match *El grupo*. This handles 3 further cases of the 18 (assuming that having matched *El grupo Roche* with *Roche*, *Productos Roche SA* will not match because it does not match with *El grupo Roche*).

If no positive partial match can be found, it is assumed that the PN is not being used to corefer to an existing referent and introduces a new referent. This takes care of 9 additional cases.

This basic PN resolution procedure, then, handles 16 of the 18 PNs (89%). That leaves 2 cases which it will not handle, *Doctor Andreu* coreferring to *Docteur Andreu* and *Productos Roche SA* coreferring to *su compañia en España*

The second problem is not all that rare. Here, the PN is being used to corefer to a referent that was initially introduced by a common noun phrase. The first step is to identify the semantic class of the PN, possibly through some independent PN classifying procedure or possibly by looking at any semantic constraints that arise from the context. In the case of *Productos Roche SA*, for instance, it might be classed as a COMPANY by some independent PN classifying procedure, say, on the basis of the *SA*, or by inspecting its context.

> … *la operación realizada entre*
> … the transaction carried out between
> *Productos Roche SA y Unión Explosivos*
> Productos Roche SA and Unión Explosivo
> *Río …*
> Río …

In example (1), the expression *la operación* is used to refer to some transaction that, as the text goes on to report, was carried out between Productos Roche and Unión Explosivos Río Tinto. If transactions are carried out by companies (general semantic knowledge), then Productos Roche and Unión Explosivos must be companies.

Having established the semantic category of Productos Roche, the next step is to establish a plausible connection between Productos Roche and an established referent of the same semantic category. That is, the procedure is now to inspect all the established referents of the category COMPANY (i.e., the Roche group, Doctor Andreu and Roche's company in Spain). We know from prior text that Roche bought Doctor Andreu and that Roche acquired Doctor Andreu through its subsidiary in Spain (epistemic knowledge). From the current text, we know that Productos Roche and Unión Explosivos were actually involved in the transaction and that Unión Explosive had been a majority shareholder but, by implication, no longer is (epistemic knowledge). Thus, Unión Explosivos appears to be the seller and, by implication, Productos Roche could be the buyer, i.e., Roche's company in Spain.

The same procedure can be used to establishing that the reference of *Doctor Andreu* is the same as that of *Docteur Andreu*: establish the semantic class of *Doctor Andreu*, inspect each existing referent of that class to see whether or not a plausible connection can be established.

4.2 Resolving Pronominals

In the sample Spanish text, there are 19 pronominal expressions, 13.9% of the explicit referring expressions or 7.3% of all references. Of these, there are 15 explicit forms and 4 ellipted forms. The 15 explicit forms include 6 possessive pronouns, 4 deictic adverbials and 5 definite articles. The 4 ellipted forms include 3 ellipted subjects of finite verbs and 1 ellipted head of a relative complementizer.

4.2.1 Explicit pronouns

The basic form-based strategy for resolving pronominal reference is to begin by inspecting in reverse order of mention those referring expressions whose forms are compatible with the morphological constraints imposed by the pronominal. This strategy is usually constrained by various syntactic heuristics such as that a non-reflexive pronoun in object position cannot corefer to the subject or a pronominal complement of a noun cannot corefer to the head (e.g., Ferrández *et al*, 1998). Such a resolution procedure will account for 4 of the 6 cases (66%) in the sample text.

To resolve the remaining cases, it is necessary to check the referent of the antecedent to see whether it is semantically compatible with the contextual function of the anaphor. So, for instance, in resolving the reference of *su* (its, his, her or their) in:

> *El beneficio neto -el mejor de su*
> The[=its] net profits the best in its
> *historia- se elevó a 641,5 millones de*
> history increased to 641.5 million
> *francos ...*
> Francs ...

the procedure is to first shuffle back through the referring expressions until a third person form is encountered. Here, the first third person referring expression is *El beneficio neto* (some company's net profits). The procedure next needs to establish through inferencing that the referent of *El beneficio neto* can serve the function of the anaphor *su*, that is, can have a history. In this case, a plausible inference cannot be established, and so the procedure moves on to consider the next most recently mentioned referent, the Roche group which is being referred to by the *El* of *El beneficio neto* which is being used to express a possessor relation. Here, on the basis of ontological knowledge about what companies can or cannot have, *su* is understood as coreferring to the referent of *El*.

4.2.2 Ellipted pronominals

For the ellipted pronominals in the sample Spanish text, syntax will have to identify such ellipted elements in order to trigger the reference resolution process. However, once identified, the basic strategy described above for explicit pronouns should apply unaltered although, unlike possessive pronouns, the morphological constraints of the anaphor must be extracted from the morphosyntactic context. So, for instance, the ellipted subject of:

> ... *cuenta* PRO *con compañías en más*
> ... has companies in more
> *de 50 países* ...
> than 50 countries ...

must be a third person singular referent (given the conjugation of the verb *cuenta*). In this case, the basic resolution procedure correctly resolves 2 of the 4 cases (50%).

For the remainder, the semantic function of the ellipted element is also extracted from context, i.e., it functions as the subject of *contar con compañías* (has companies). Thus, in the example above, it must be something that can own companies. Among the third singular candidate expressions, in reverse order of mention: *el diagnóstico* (diagnosis), *la comercialización* (the marketing), *la producción* (the manufacture), *el desarrollo* (the development), *Basilea (Suiza)* (Basel, Switzerland) and *sede central* (home office), none are potential owners of companies. The next most remote referring expression to be inspected is *el grupo Roche* (the Roche Group) which, as it turns out, is something that can own companies and, therefore, the PRO is identified as coreferring to the same referent.

4.2.3 Deictic Elements

Deictic elements, such as the adverbs *hoy* (today), *aquí* (here), *ahora* (now), and so on, are resolved directly to properties of the utterance context: the day of utterance, the place of utterance, the time of utterance, and so on. There were 4 such pronouns in the sample text, 3 referring (*hoy*, *aquí*, *ahora*") and 1 coreferring (*hoy*). It should be noted that all these elements are, in fact, coreferent with implicit temporal and spatial references of various finite verbs that are used to report certain events or states of affairs.

4.2.4 Articles with pronominal force

Perhaps the most contentious of the pronominal elements to be discussed here are the definite articles of noun phrases which can be contextually interpreted as having the force of possessive adjectives. There are 6 examples of this in the sample text. However, these are by no means all the definite articles found and, in addition, of these six example, four were translated as possessive adjectives, one as a definite article, *the*, and one was omitted altogether in translation. In other words, not only is an ambiguity introduced for a very common lexical item, but even when resolved in favor of the possessor interpretation, it may not be translated as a possessive adjective. On the other hand, the major reason for assuming a distinction is that it is very important to establish such relationships in order to understand or translate a document. For instance, if it is not established that the cash flow referred to in:

> *El "cash flow" se incrementó en un 21*
> **its** cash flow increased by about 21
> *por ciento ...*
> per cent ...

is that of company X, then it will be impossible to determine whether the cash flow referred to later in the text in:

> *... y el "cash flow" de ...*
> ... and **its** cash flow ...

is that of company X or of some other company. That means that such information will not only be unavailable for use during translation (e.g., selecting a possessive adjective in the target language) but for any other purpose that might come along (e.g., information extraction).

In any case, the procedure for resolving the reference of definite articles is the basic pronominal resolution procedure except that no morphological constraints can be placed on the antecedent expression. Still, syntactic constraints on the antecedent may be applied. Following this strategy, 5 of the 6 cases (83%) are resolved correctly although it is not clear whether it leads to false positives. Otherwise, potential referents are considered until one is found which can serve the appropriate possessor function. Since a positive connection must be inferred, the likelihood of false positives is greatly decreased.

4.3 Resolving Common Noun Phrases

If proper noun phrases and pronominals were the only type of referring expressions, form-based resolution techniques might prove sufficient. They account, however, for only 23.6% of the explicit referring expression or 12.4% of all references. It is for the resolution of common noun phrases, clauses and implicit references that an interlingual-based procedure will eventually prove necessary.

There are 76 common noun expressions in sample Spanish text. Of these 42 are definite noun phrases (32 referring, 10 coreferring), 9 are indefinite noun phrases (all 9 referring), and 25 are noun phrases having no article (all 25 referring).

Since none of the indefinite noun phrases or the noun phrases without articles are used to corefer, an initial basic resolution strategy for common noun phrases begins by inspecting the form of the referring expression. If it is an indefinite noun phrase or noun phrase without article, it is assumed to refer to something new and a new referent is added to the referents in the domain of discourse. That successfully resolves the reference of 34 of the 76 common noun phrases (45%) in the text.

Of the 42 definite noun phrases, 36 are used to refer (or corefer) to specific individuals or stuff or are inferrably unique given a general knowledge of individuals, stuff or situations that were being discussed. Of these 36, 24 were used to refer to particular individuals or stuff, 8 to processes, 2 to particular groups of objects and 2 to logically unique objects. Of the remaining 6 definite noun phrases, 4 were used to refer to portions (percentages) of stuff and 2 were used to refer to generic classes.

In regard to the resolution of definite noun phrases, then, the basic strategy is to first identify whether the expression is being used to refer to specific individuals or stuff, to portions (percentages) of stuff or, if possible, to a generic class. So, for instance, given the expression *el 7,4 por ciento* in:

> *la rentabilidad sobre las ventas aumentó*
> its profit over sales increased
> *del 6,3 al **7,4 por ciento**.*
> from 6.3 to 7.4 per cent

it is sufficient to identify that the expression is being used to refer to a percentage (of sales) in order to assume the the expression is used to refer to something new. Identifying a generic reference on the basis of form is less obvious but, in any case, this will resolve 4 to 6 of the 42 cases.

Second, for the 36 definite noun phrases used to refer to a particular individual or stuff, the basic procedure is to match the head noun expression against each of the common noun expressions that have been used previously. If one is found, the complement expressions are then matched. If these are compatible, the NP under consideration is assumed to corefer to the referent of the matching NP. This will successfully resolve 27 of the 36 cases (75%) but leaves 9 incorrectly resolved, 5 setting up new referents when in fact they are coreferring and 4 false positive cases of coreference.

The third step, then, is to inspect each referent for semantic compatibility. Semantic information is established on the basis of the expression's form and context. For instance, in looking for a possible referent for *la rentabilidad sobre las ventas* (its profit over sales ratio) above, it is first necessary to establish that the potential referent must be some measurement of financial performance which has increased during some particular period of time for some particular company.

At the time a reference for *la rentabilidad sobre las ventas* is sought, there are some 98 existing referents (53 objects, 34 events, 11 implicit objects). The more recent of these include “641.5 million Swiss Francs”, “Roche Group's net profit”, “Roche Group's pharmaceutical division”, “41% of Roche Group's total sales”, “8.69 billion Swiss Francs”, “Roche Group's total sales”, “1988” and so on. Of these, only “Roche Group's net profit” and “Roche Group’s total sales” are possible measurements of a company's financial performance. However, both these measurements should be ontologically distinct from a company's profit over sales as well as from each other. Thus, they fail to satisfy the semantic requirements of a potential referent for the expression. In the end, no semantically appropriate referent will be found among the pool of existing referents and so a new referent is introduced to the pool.

If an existing referent meets the informational constraints on potential referent of the expression being processed, the expression is assumed to refer to that referent. If no existing referent satisfies those constraints, the expression is assumed to refer to something new and a new referent is added to the pool of referents.

4.2.4 Resolving Clauses

Noun phrases, of course, are not the only type of constituent that is used to refer to things in the world. Clauses may be used to refer (or corefer) to particular events or states-of-affairs or to classes of events or states-of-affairs. These may be finite (main, relative, complement or adverbial clauses), participials (present or passive), infinitival or absolutive. Of the 45 events referred to in the text, there were 3 events that were referred to on more than one occasion. The first, the purchasing of Doctor Andreu, was coreferred to 4 times. But of these, only one was by way of a (finite) clause. The other three were all by way of NPs, 2 explicit and 1 implicit. The second, the announcing of the purchase, was coreferred to only once by way of a NP. The third, Roche’s investing in R&D, was coreferred to once by way of an implicit pro-verb introduced for syntactic reasons in the context of a parallel, conjoined structure.

The only form-based resolution procedure for resolving clausal reference would be to look for prior verbs having the same form and then inspecting the complements for contradictions. This procedure might be extended by inspecting prior verb forms that are related by, say, Spanish WordNet (Rodríguez, 1998) or an on-line Spanish thesaurus (if any should exist). However, this extension could also open the door to many false positives.

Such an approach might possibly resolve as many as 26 of the 27 cases of clausal references to events or states of affairs correctly. In any case, an IL-based approach has the advantage of having the events mentioned in prior text already represented formally and in a language neutral form. Thus, the need for additional on-line resources for each language is assuaged.

4.2.5 Resolving Implicit References

As mentioned, there were 45 events or states-of-affairs referred to in the sample Spanish text which introduce an additional 30 implied referents (5 times, 22 places and 3 actors). These events, and the implicit referents they introduce, need to be identified for successfully carrying out the coreference task. They may act as referents that are later referred to in the text or they may serve to assist in constraining or establishing coreference between later expressions and other existing referents. For instance, of the 22 implied locations, 6 are later referred to in the text and, of the 5 implied times, 3 are later referred to.

Clearly, there is no obvious form-based resolution procedure for such elements, since they have no explicit form. Thus, in order to resolve these implicitly introduced referents, the basic procedure is to treat them as pronominals. That is, every event or state-of-affairs in the TMR has an implicit "at that time" and "at that place" associated with it which has to be resolved as part of the reference task. Beyond the fact that the potential referents must times and locations respectively, any further constraints will have to be derived from what is known about the event and about its relative (temporal or local) status with respect to the other events which have previously been mentioned. A primary source of information for dealing with such issues will be scripts (Schank & Ableson, 1977). Similarly, the identification of implied actors will be dependent on the ontological (not lexical) definitions of the class of event or state-of-affairs referred to and any additional information that may be extracted from the particular events or states-of-affairs that have been previously mentioned.

Given the informational constraints gathered, the procedure is then to inspect referents of like type (time, location, actor type) in reverse order of mention until one is found which is compatible with those additional informational constraints.

Conclusion

The advantages of the interlingual approach to reference resolution include the following:

- only expressions related to actual referents are processed for coreference (i.e., no pleonastic pronouns, no clitic pronouns, no relative pronouns, etc.),
- implicit as well as explicit referents are processes for coreference,
- knowledge-based inferencing (both ontological and epistemic) is available for resolving (many) problematical cases,
- ontologically connected actors, say, the different participants in a sequence of events making up a script, can be used to establish coreference,
- texts in different languages can be processed in the same way,
- all form-based procedures either are or can be implemented in any case.

The central disadvantages are:

- some surface text level ordering information is lost in the TMR,
- discourse-structural information may be lost in the TMR,
- the need for a large and sophisticated knowledge sources,
- the need for sound and appropriately-directed inferencing..

As a result of the loss of ordering information, strict recency-based resolution procedures cannot be implemented. The referents in the domain are not ordered in terms of when they were introduced. The processing of the different arguments in f-structure does not necessarily correspond to the surface sequence of their mention. This "defect" could possibly be overcome by simply indexing each new the TMR object with the prior index plus one (assuming the indexes are integers). The tacit assumption is that, at the level of the clause, first the predicate is processed and then the arguments are processed in left to right order as they appear in f-structure.

As for the lack of information about the discourse structure, it may be the case that this is a defect of the TMR representation system. That is to say, it is not unreasonable to assume that the larger organisational aspects of a text, the topics, their order of presentation, the structure of the argumentation, etc., should in fact be captured in any adequate representation of the text. It has been, however, the goal of TMR to focus on capturing the information content exclusively and not on how the information is presented.

References

Bagga, A., and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector-Space Model. *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics and the 17th Conference on Computational Linguistics*, Montreal, Canada.

Carlson, L., and S. Nirenburg. 1990. *World Modeling for NLP*. Technical Report 90-121. Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.

Ferrández, A., M. Palomar, and L. Moreno. 1998. Anaphora resolution in unrestricted texts with partial parsing. *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics and the 17th Conference on Computational Linguistics*, Montreal, Canada.

DeJong, G. 1979. *Skimming stories in real time: an experiment in integrated understanding*. Ph.D. dissertation. Research report #158, Dept. of Computer Science, Yale University, New Haven, CT.

Grosz, B., A. Joshi, and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2): 203-225.

Hirschman, L., P. Robinson, J. Burger, and M. Vilain. 1998. Automating Coreference: The Role of Annotated Training Data. *Proceedings of AAAI 98 Spring Symposium on Applying Machine Learning to Discourse Processing.*

Hobbs, J. 1979. Coherence and Coreference. *Cognitive Science* 3: 67-90.

Lappin, S., and H. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4), 535-561.

Mahesh, K., and S. Nirenburg. 1995. A Situated Ontology for Practical NLP. *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence*. Montréal, Canada.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference, April 1998.*

Onyshkevych, B., and S. Nirenburg. (1994). *The Lexicon in the Scheme of KBMT Things.* Memoranda in Cognitive and Computer Science: MCCS 94-277, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Popescu-Belis, A. 1998. How Corpora with Anotated Coreference Links Improve Reference Resolution. In Rubio, A., N. Gallardo, R. Castro & A. Tejada (eds.) *Proceedings of the first International Conference on Language Resource and Evaluation,* pp. 567-571. Granada, Spain, May 1998.

Rodríguez, H., S. Climent, P. Vossen, L. Bloksma, W.Peters, A. Alonge, F. Bertagna, and A. Roventini. 1998. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. *Computers and the Humanities*, 32(2/3): 117-152

Schank, R., and R. Abelson. 1977. *Scripts, plans, goals, and understanding.* Lawrence Erlbaum Associates: Hillsdale, NJ.

Sundheim, B., and R. Grishman (eds.). 1995. *Sixth Message Understanding Conference (MUC-6).* San Francisco: Morgan Kaufman.

White, J., T. O'Connell and F. O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches, in Technology Partnerships for Crossing the Language Barrier: *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, MD, pp. 193-205.

Appendix

Spanish Evaluation Text 93-1

Roche Compra Docteur Andreu

El grupo Roche, a través de su compañía en España, adquirió el laboratorio farmacéutico Doctor Andreu, se informó hoy aquí.

La comunicación oficial no precisó el monto de la operación realizada entre Productos Roche SA y Unión Explosivos Río Tinto SA, hasta ahora mayoritaria en el accionariado.

Fuentes financieras consultadas cifraron la operación en unos 10.000 millones de pesetas. Según el acuerdo firmado hoy en Madrid, los productos del Doctor Andreu continuarán siendo producidos y comercializados con el mismo nombre. Doctor Andreu, cuya fama la obtuvo a partir de las "pastillitas" para la tos, está bien introducido en las áreas de cardiología, reumatología y especialidades publicitarias.

Las actividades del grupo Roche, con sede central en Basilea (Suiza), incluyen el desarrollo, la producción y la comercialización de medicamentos, productos para el diagnóstico, así como de vitaminas y productos químicos.

A nivel mundial, cuenta con compañías en más de 50 países con casi 50.000 empleados. Doctor Andreu es una compañía farmacéutica dedicada a la producción y comercialización de fármacos y productos veterinarios. Con sede en Barcelona, cuenta con más de 400 empleados.

En el ejercicio pasado facturó unos 3.490 millones de pesetas.

En 1988, el Grupo Roche alcanzó unas ventas totales de 8.690 millones de francos suizos, de las que aproximadamente un 41 por ciento correspondieron a su división farmacéutica. El beneficio neto -el mejor de su historia- se elevó a 641,5 millones de francos suizos y la rentabilidad sobre las ventas aumentó del 6,3 al 7,4 por ciento.

El "cash flow" se incrementó en un 21 por ciento, alcanzando 1.179 millones de francos o el 14 por ciento de las ventas del grupo.

Las inversiones en investigación y desarrollo (I+D) fueron de 1.210 millones de francos suizos, el 14 por ciento del total de sus ventas.

Productos Roche cuenta con una plantilla de 600 personas y alcanzó unas ventas totales de 9.747 millones de pesetas, un 12,5 por ciento superiores al año 1987.

Sus beneficios fueron de 218 millones y el "cash flow" de 356 millones. Las inversiones realizadas totalizaron 223 millones de pesetas.

English Translation of Text 93-1

Roche Buys Docteur Andreu

The Roche Group acquired the pharmaceutical laboratory Doctor Andreu through its company in Spain, it was announced here today.

The official announcement did not specify the exact amount of the transaction which took place between Productos Roche SA and Unión Explosivos Río Tinto SA, which until now had held the mayority of the stock.

Financial sources consulted estimate the transaction value at around 10 million pesetas. According to the agreement signed today in Madrid, Doctor Andreu's products will continue to be produced and marketed under the same name. Doctor Andreu, which became well known for its cough drops, is well established in the areas of cardiology, rheumatology, and advertising specialties.

Activities of the Roche Group, headquartered in Basel (Switzerland), include the development, production, and marketing of drugs, diagnostic products, as well as vitamins and chemical products.

The Roche Group has subsidiaries in more than 50 countries and almost 50,000 employees worldwide. Doctor Andreu is a pharmaceutical company which produces and markets medicines and veterinary products. Headquartered in Barcelona, it has more than 400 employees.

In the last fiscal year its sales totaled some 3.49 billion pesetas.

In 1988, the Roche Group's total sales reached 8.69 billion Swiss francs, of which approximately 41 per cent corresponded to its pharmaceutical division. Its net profits--the best in its history--went up to 641.5 million Swiss francs and the profitability over sales increased from 6.3 to 7.4 per cent.

Its cash flow increased by 21 per cent, to 1.179 billion francs or 14 per cent of the group's sales.

Investments in research and development (R&D) totaled 1.210 billion Swiss francs or 14 per cent of total sales.

Productos Roche employs 600 people and its total sales reached 9.747 billion pesetas, 12.5 per cent more than in 1987.

Its profits totaled 218 million and its cash flow 356 million. Actual investments totaled 223 million pesetas.